Varying correlation coefficients cannot account for uncertainty about dependence, but there are comprehensive methods to do so

Scott Ferson, Applied Biomathematics, 100 North Country Road, Setauket, New York 11733, 631-751-4350, fax –3435, scott@ramas.com

Extended Abstract. Although risk analyses often still assume independence among input variables as a matter of mathematical convenience, most analysts recognize that inter-variable dependencies can sometimes have a substantial impact on computational results. In the face of epistemic uncertainty about dependencies, analysts occasionally employ a sensitivity study in which the correlation coefficient is varied between plausible values. This strategy is insufficient to explore the possible range of results however, as will be shown by simple examples. Fortunately, comprehensive bounds on convolutions of probability distributions (or even bounds thereon) can be obtained using simple formulas that are computationally cheaper than Monte Carlo methods. We review the use of these formulas in the cases of variously restricting assumptions about dependence, from no assumption at all, to specified sign of the dependence, to a particular dependence function.

Some analysts argue that it is best to reduce any problem involving dependent variables into one with only independent variables. This changes the problem of statistically representing dependent variables into a modeling problem of reproducing the functional or mechanistic relationships that induce the dependence. It is not sufficient to transform the model into one in terms of uncorrelated variables; they must be statistically *independent* variables. Of course, this functional modeling approach could entail considerable effort far beyond the scope of the immediate assessment. The extra modeling effort required by this strategy may not be workable in many situations. For instance, a dam safety engineer worried about computing risks to a water control structure from hydrological factors influenced by weather patterns would need to model various meteorological and even climatological phenomena. At some point, the analytical demands of a functional modeling approach will likely become prohibitive.

There are three other approaches to the problem of accounting for dependence among variables: (*i*) assume a particular dependence function, (*ii*) make no assumptions about dependence, and (*iii*) relax assumptions to a partial specification of dependence. In the first approach, one must assume a particular dependence function among the variables. Assuming independence is of course a special case of this approach. Another special case is assuming perfect dependence among variables such that each variable is almost surely a monotonically increasing function of the other. In general, the dependence function is specified as some copula (Nelsen 1999). A copula is the function that characterizes how the marginal distributions are knitted together to form the joint distribution. In the two-dimensional case, a copula is just a bivariate distribution function from the unit square onto the unit interval that has uniform marginals. A bivariate distribution function $F(x, y)$ can be expressed in terms of the copula C as $C(F(x), G(y))$ where $F(x)$ and $G(y)$ are its marginal distribution functions. The dependence function could be specified by selecting a copula from a parameterized family of copulas such as

the Frank, Mardia, normal or Clayton families (Nelsen 1999; Joe 1997). It could also be specified with an empirical copula, which is an analog for dependence of an empirical distribution function.

In risk analyses, distributions characterizing random variables are convolved together to estimate arithmetic functions (such as sums, products, differences, quotients, etc.) of the random variables. For instance, if X and Y are random variables with distributions F and G respectively, the distribution of the sum $Z = X+Y$ can be obtained with the Lesbesgue-Steiltjes integral

$$\sigma_{+,C}(F,G)(z) = \int_{x+y<z} dC\big(F(x), G(y)\big)$$

which always exists. This formulation includes the independence case where $C(u,v) = uv$. Similar formulas are available to compute distributions of products, differences, quotients, etc. We describe a straightforward numerical procedure to compute σ given discretizations for the marginal distributions F and G and an arbitrary copula C. The numerical methods extend easily to other arithmetic operations. Note that this approach can handle arbitrarily complicated dependence between the input variables. This makes the approach significantly more general than methods implemented in common risk analysis software packages which model correlations but not dependencies in general.

The second approach to accounting for dependence is to make *no assumptions whatever* about the dependence between variables. In this approach, bounds on the distribution of an arithmetic function can be computed directly using infimal and supremal convolution of the marginal distribution of the addends. For example, if X and Y are random variables with marginal distributions F and G respectively, then the bounds on the distribution of $Z = X+Y$ are

$$\left[\sup_{z=x+y} \max\big(F(x)+G(y)-1, 0\big), \inf_{z=x+y} \min\big(F(x)+G(y), 0\big)\right]$$

where the supremum gives the left bound on the distribution (i.e., the upper bound on the cumulative probability associated with any value of the sum z), and the infimum gives the right bound on the distribution function (the lower bound on the value of the cumulative probability). These bounds satisfy a problem originally posed by Kolmogorov of finding bounds on the distribution of a sum given only distributions of the addends. Kolmogorov's problem was solved by Makarov (1981) and Frank et al. (1987). Analogous bounds on distributions of products, differences, quotients, etc., can likewise be obtained from similar supremal and infimal convolutions of the marginal distributions of the factors, etc. Williamson and Downs (1990) described convenient numerical algorithms to compute these bounds in a way that accounts for discretization error introduced by encoding the marginal distribution with a finite computer representation. With their algorithms, the bounding convolutions are generally much less expensive than ordinary convolution computed via Monte Carlo simulation. The bounds obtained by the

supremal and infimal convolutions are guaranteed to rigorously enclose all distributions that could arise for the sum (or product, etc.), no matter what dependence there may be between the addends (or factors, etc.). Furthermore, these bounds are also best possible, that is, they are as tight as can be justified without any knowledge about the dependence. The breadth between the bounds characterizes the specificity of the answer under the relaxed dependence assumption. It is interesting that these bounds *cannot* be obtained with the standard σ-convolution described above such as by varying the correlation between +1 and –1. Even varying the dependence function C between perfect dependence (maximal correlation and comonotonicity) and opposite dependence (minimal correlation and countermonotonicity) will generally underestimate the breadth of the bounds. The difference is due to nonlinear dependencies which are ignored by merely varying correlations between extreme values. This approach can be combined with independence assumptions, so that some variables are assumed to be independent and no assumptions are made about the dependence between other variables.

The third approach to account for dependence in risk assessments is to make some qualitative or quantitative assumptions about the dependence function that partially specify the copula. For instance, a promising approach to tighten risk calculations is to make use of information about the *sign* of the dependence between the variables. The most common notion of sign dependence is positive quadrant dependence (PQD). Random variables X and Y with distribution functions F and G whose joint distribution is H are PQD if $H(x, y) \geq F(x)\ G(y)$ for all x and y, so that if the probability that the random variables are both small (or large) is at least as great as if they were independent. There are several conditions that imply variables will be PQD, including when each is a stochastically increasing function of the other, i.e., $\mathrm{P}(Y > y \mid X{=}x)$ is a non-decreasing function of x for all y, and $\mathrm{P}(X > x \mid Y{=}y)$ is a non-decreasing function of y for all x. Positive quadrant dependence implies non-negative Pearson, Spearman and Kendall correlations, although the mere observation that a correlation is positive does not imply the variables are PQD. This idea has been used in many statistical and engineering settings, and seems to capture one sense analysts have in mind when they use the phrase 'positively depends'.

Risk assessments can make use of assumptions about the sign of the dependence among variables with easy-to-compute convolutions. For example, bounds for a sum of PQD variables whose marginals are F and G are

$$\left[\sup_{z=x+y} \left(F(x)G(y)\right), \inf_{z=x+y} \left(1-(1-F(x))(1-G(y))\right) \right].$$

These bounds are similar to the supremal and infimal convolutions in the sense that they are guaranteed to bound the distribution function of the sum and are the tightest possible such bounds given only the marginal distributions F and G and the positivity of their dependence. Note that these formulas give bounds that are *not* the same as an envelope of the perfect and independent convolutions (which would be narrower). There are similar formulas for the other arithmetic operations, as well as complementary formulas

that assume negative quadrant dependence (variables X and Y are negatively quadrant dependent if X and $-Y$ are positively quadrant dependent). The intersection of the convolution bounds for positive and negative dependencies is *not* the same as the bounds obtained under independence.

One could also make a *quantitative* assumption about dependence such as that the correlation coefficient has a particular magnitude. In such cases, convolutions between distributions can be computed using mathematical programming, although it turns out that specifying only the correlation often provides very little improvement in the specificity of the result. For example, assuming that random variables X and Y are uncorrelated (that is, have Pearson correlation coefficient equal to zero) produces almost no improvement over the bounds obtained by the supremal and infimal convolutions.

The three approaches described above give analysts considerable flexibility to account for knowledge and uncertainty about correlations and dependencies. By making more assumptions, one can increase specificity of the answers that can be obtained. In a sensitivity analysis, of course, an analyst often desires to relax his assumptions and explore how the results might vary in consequence. It is possible to mix strategies so that one could posit independence among some variables, assume particular copulas for some variables, and make limited or no assumptions about the dependence among other variables. This allows an analyst to obtain a sensitivity analysis that reflects which is well known about dependencies and what is in contention about them.

Acknowledgments. This work was supported under contract 19094 with Sandia National Laboratories. It benefited from discussions with Roger Nelsen, Daniel Berleant, Jianzhong Zhang, and Roger Cooke.

References

Frank, M.J., R.B. Nelsen, and B. Schweizer. 1987. Best-possible bounds for the distribution of a sum—a problem of Kolmogorov. *Probability Theory and Related Fields* 74: 199-211.

Joe, H. 1997. *Multivariate Models and Dependence Concepts*. Chapman & Hall/CRC, Boca Raton.

Makarov, G.D. 1981. Estimates for the distribution function of a sum of two random variables when the marginal distributions are fixed. *Theory of Probability and its Applications* 26.

Nelsen, R.B. 1999. *An Introduction to Copulas*. Springer-Verlag, New York.

Williamson, R.C. and T. Downs. 1990. Probabilistic arithmetic I: Numerical methods for calculating convolutions and dependency bounds. *International Journal of Approximate Reasoning* 4: 89-158.